『学習院大学　経済論集』第46巻　第1号（2009年4月）

混合分布問題
――その基礎からカーネル降下法まで―― Part 1

金田　尚久、新居　玄武

１．はじめに

統計学で，通常用いられる確率分布は，一様分布を除けば，全て，一つの峰を持った分布である。ところが，諸科学の応用においては，二つ以上の峰を持った分布を考えねばならない場合がある。このようなとき，一番単純なモデルは，一つの峰を持った分布を必要な数だけそろえ，そのそれぞれにウェイトをかけて，たし合わせたものである。

$$f=\sum_{i=1}^{n} w_i f_i \qquad (w_i>0,\quad w_1+w_2+\cdots\cdots+w_n=1)$$

ここで，一つの峰を持った分布 f_i の個数nがわかっているなら，それぞれの f_i に含まれるパラメターの推定が問題になる。もしnがわかっていないのなら，nをどうやって推定するかが，問題として付け加わる。このような形で与えられた問題を混合分布問題と呼ぶ。19世紀の終わりにカール・ピアソンが，生物学の問題に関連して考察して以来，混合分布問題は，長い研究史を持つ。しかし，その歩みは，決して平坦ではなかった。まず第一に，ピアソンの最初の解法が余りに難解で，普遍性があるかどうかも疑わしかった。この方法では，問題は，積率法を通して，9次の代数方程式を解くことに帰着される。これは，コンピュータのなかった時代には，余りにも，大きな労力である。そこで，もっと計算の負担が少く，見通しのよい解法に改良する試みが現れた。しかし，これらの努力も，推定値の格別な向上には，つながらなかった。この後，研究は長い停滞の時期を迎える。新しい動きが見えたのは，コンピュータが普及し始める1960年代である。これ以後，この問題のために提案されたアルゴリズムは，コンピュータの使用を前提とし，くり返し計算に全面的に依存している。その中で代表的なのは，EMアルゴリズムとマルコフ連鎖モンテカルロ法（MCMC）である。（いずれも，後章で詳しく説明する。）さて，初めに述べたように，混合分布問題には，nが与えられている場合と，推定しなければならない場合と，二種類ある。これまでの様々な応用例から見る限り，EM とMCMCは，nが固定されている場合には，まずまずの成果が挙がっている。しかし，nを推定する問題には，充分でない。Scott と Szewczyk (SS) は，2001年に，この2つのアルゴリズムと全く異なる，新しいアルゴリズムを開発した[11]（文献はPart 2の最後にまとめて掲げる）。これは，1次元正規混合分布における，nの推定を目的としたアルゴリズムである。本論文の共著者の一人（金田）は，この方法に関心を持ち，そこで使われる新しいアイディアの解明と，同じアルゴリズムの他の混合分布への拡張を課題として，Ph.D.論文を執筆した[6]。本論文の

トピックの一つは，そこで得られた成果を，初めて日本語で紹介することである。本論文においては，混合分布問題の初歩から，最新の成果までのバランスの取れた解説を目標としている。次章でも述べるように，混合分布問題は，現在，理工学の幅広い分野で，熱烈な関心を持たれている研究テーマである。その簡明で一般的な問題設定から，我々は，社会科学方面にも大きな潜在的応用可能性を期待し得る。しかし，この問題は，現在までのところ，最大の応用分野が画像処理であることも手伝ってか，「先端工学」的なイメージが強い。社会科学への応用の試みが乏しいのも，その辺のところに理由があろう。そこで我々は，経済学部で通常教えられる，数学の知識を前提として，社会科学者のための，この問題のチュートリアルを提供することは，極めて望ましいと考えた。以下，このプランに沿って，混合分布問題の意味・EMアルゴリズム・MCMC・SSのアルゴリズム・金田による研究の順に解説する。

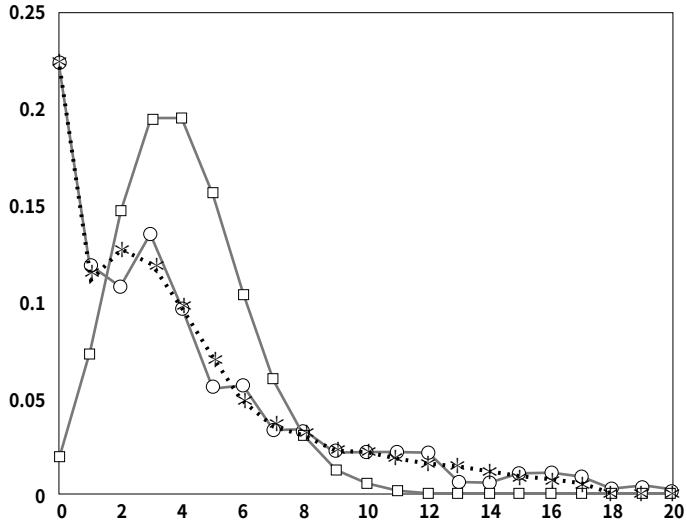
２．混合分布で，何がわかるのか？

1980年代半ばに出版された，Titterington,Smith,and Makov [12] は，それまでの混合分布研究の総まとめとして，意義深く，2000年にMcLachlan and Peel[8] が出るまで，この分野の標準的モノグラフであった。そのp.16～p.21には，Pearson以来その当時までの，応用例が，分野別の表としてまとめられている。これを見ると，医学・心理学・動物学・植物学・地質学・工学・ORなど，幅広い分野で，すでに混合分布が応用されていたことがわかる。応用分野は，その後も広がっているが，人文・社会科学の中で，比較的早くから混合分布の応用に熱心だったのは，マーケティングと心理学（精神医学）である。この二つの分野から，それぞれ一つずつ例を取って，混合分布を応用する動機を見てみよう。

例１．キャンディーの購買数に関する研究

DillonとKumar[2] は，456人の消費者が，一週間に買ったキャンディーの数を調べた。このデータ全体を確率分布として把握するために，キャンディーの箱の数xを横軸に，
（x箱のキャンディーを買った人の数）/ 456 を縦軸にとる。この平面上にデータをプロットしたのが，下の実線と丸で表わされたグラフである。

図１　キャンディーの購買数のデータと二つの推定法



実線と丸：データ
実線と四角：一つのPoisson 分布によるあてはめ
点線とアステリスク：6クラスター・モデル

このデータの形状は示唆深い。一定時間の顧客の数は，Poisson分布によって，よく近似される。

$$f(x,\mu)=\frac{\mu^x e^{-\mu}}{x!}$$

平均$\mu<1$のとき，この分布は単調減少である。データはx=0で最大値を取り，全体に右下がりの傾向を示している。このことは，この期間中に，1箱もキャンディーを買わなかった消費者が，多かったことを意味する。ところが，標準的な最尤法でパラメターを推定すると，

$\hat{\mu}=\sum_{i=1}^{N}x_i \Big/ N=3.991 \quad (N=456)$ となり，平均は4に近くなり，一つの峰を持ったPoisson分布

が現れてしまう。このことから，$\mu<1$のように見えるのは，見かけ上に過ぎず，この経験分布の右側のtailには，相当に確率質量が集積していることがわかる。さらに，小さな上がり下がりを追っていくと，このデータにはx=3,6,8,12及び15から17にかけてのところに，小さな峰がある。このような特徴をつかむには，いくつかのPoisson分布の重ね合わせと考えるのが適当であろう。DillonとKumarは，3章で説明する，EMアルゴリズムによって，混合分布をあてはめた。彼らは6つのPoisson分布の混合が最適と結論している。行列の第1列に平均，第2列にウェイトを書いて，6つのクラスターは，以下のように表せる。

$$\begin{bmatrix} 0 & 0.1555 \\ 1.079 & 0.1342 \\ 3.209 & 0.4737 \\ 7.534 & 0.1277 \\ 11.127 & 0.03 \\ 13.6 & 0.0789 \end{bmatrix}$$

図1の点線とアステリスクのグラフは，この6つのクラスターを，ウェイトを付けて足し上げたものである。最尤法であてはめたPoisson分布が，データの形状からかけ離れているのと比べて，この6クラスター・モデルは，はるかに良くなっている。しかし，x=3の大きな峰がとらえられず，x=2のところに峰ができてしまっていること，6つのクラスターを使っているにもかかわらず，その他の小さな峰は，ほとんどとらえられていないこと，などを考えれば，まだ改善の余地はありそうである。

例２．統合失調症の発症時期に関する研究

統合失調症（精神分裂病）は，多くの病気が征服された今日でも，解明の困難な病気である。ガンについては，発生機構に不明な点が多いにせよ，有効な対症療法は，数多く開発された。これに対して，統合失調症の対症療法には，一時的に患者の気分を軽快にする手段しかなく，その発生機構にいたっては，ほとんど何もわかっていない。しかし，19世紀からのデータの蓄積によって，この病気の多くの側面に，他の病気には見られない規則性が存在することが，わかってきた。それらの規則性には，データを取りまとめる際の判断基準にあいまいな点があり，そこを明らかにしないと，夾雑物が混じったままでしか提示できないといわれている。しかし，その存在に関しては，多くの精神科医が認めているので，将来の病気の本質解明に役立つ知見として，期待されている。そのような規則性の一つは，発症時期に関して，男女間に，共通点と相違点があることである。Lewine[7]によれば，統合失調症には早い時期（青年期）に発症する群と，遅い時期（壮年期）に発症する群と，二つのタイプがある。前者は主に男性，後者は主に女性の病気である。Everitt [3] は，この仮説をデータによって検証した。図2aと2bでは，男女それぞれの発症時期のデータが棒グラフで表わされている。しかし，グラフを見やすくするために，横軸に年齢ではなく，log（年齢）を取り，これを等分の階級に分けて，それぞれの階級に落ちた観測値の数を，男性または女性の総数で割った値が，棒グラフの高さである。（ただし，図2a,bの説明参照）即ち，男女それぞれの経験分布が読み取れるように，調整されたグラフである。

図２ａ　男性の発症時期

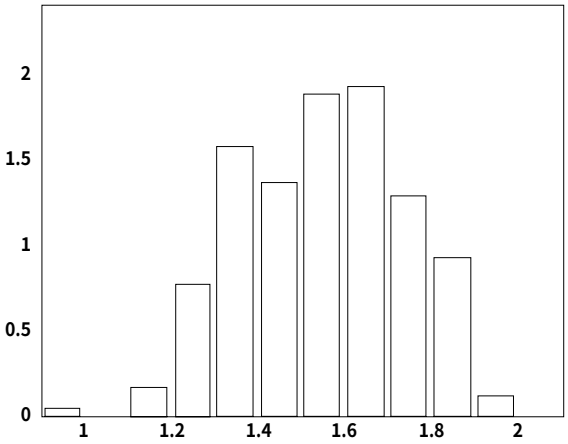
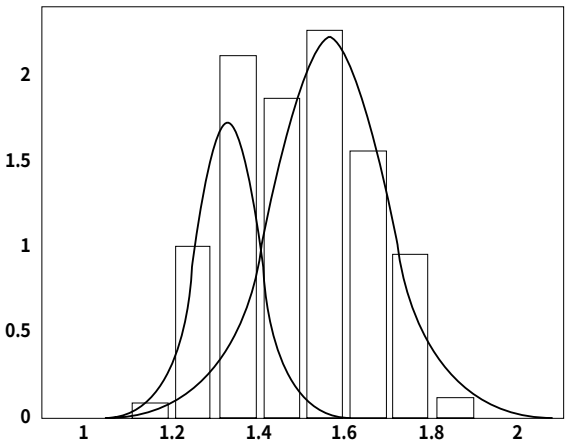


図２ｂ　女性の発症時期と２クラスター・モデル



縦軸のメモリは図2bで推定された確率分布にふさわしく取ってある。この分布との比較を容易にするために，棒グラフは，男女とも縦方向に約10倍に引き伸ばしてある。log（年齢）と年齢の対応関係は，以下の表を参照。

log（年齢）	1	1.2	1.4	1.6	1.8	2
年齢	10	15.85	25.12	39.81	63.10	100

これを見ると，確かに青年期と壮年期に高い山があり，その間はくぼんでいる。しかし，Lewineが言うような，早発型は主に男性の病気であり，後発型は主に女性の病気であるという違いは見出せるだろうか？　むしろ，男性では高齢で発症する者がいるが，女性ではそのようなケースが少いということが，大きな違いではないかと思われる。そもそも，Lewineは

“ the early onset, typical schizophrenia is largely a disorder in men, and late onset, atypical schizophrenia is largely a disorder in women “

と言うのだが，“ largely a disorder in men ”とか“ largely a disorder in women ”とは，何を意味するのだろう？　男性の分布の中で7:3女性の分布の中で4:6のように，早発性の男性と後発性の女性は，両方とも50％を越えているという意味だろうか？　それとも，男性の分布の中で4:6女性の分布の中で3:7のように，男女とも，早発・後発のいずれかに50％以上偏っている場合を許すのだろうか？　このようにLewineの主張にはアイマイな点があるが，Everitt は以下のように議論を進める。まず，男女それぞれのデータに2クラスター正規分布を当てはめ，次の結果を得た。

表１．Everitt による2クラスター・モデルの当てはめ

	μ	σ	*w*
男性早発	**1.5236**	**0.1734**	**0.919**
男性後発	**1.8267**	**0.0371**	**0.081**
女性早発	**1.3337**	**0.0731**	**0.315**
女性後発	**1.5663**	**0.1229**	**0.685**

当てはめの方法はEMアルゴリズムである。後述するように，EMは，全パラメターに適当な初期値を与えた後，繰り返し計算によって，これらを最適な値に近づけていく，アルゴリズムである。Everittは種々の理由から，Pearsonの積率法が，この初期値の決定に利用できるとしている。しかし，これは今日，パッケージなどで一般に用いられている方法ではない。コンピュータの導入によって，Pearsonの時代よりも，計算の負担は楽に乗りこえられるようになったが，3クラスター以上の場合，多次元の場合に，この方法の拡張は，尚，適当でないからである。とにかく，このようにして得られた2クラスター・モデルの比較の対象として，Everittは，1クラスター・モデル（1つの正規分布）を，男女それぞれの分布にあてはめる（推定法と推定値は示されていない）。次に，2クラスター・モデルが本当に必要かどうか知るために，2クラスター・モデル対1クラスター・モデルの検定を行う。混合分布問題は，推定だけでも充分難しいので，検定はさらに難しい。この種の検定も昔から提案されているが，これには，大きな問題が指摘されている。しかし，Everittは，この検定の結果として，女性については2クラスター・モデルが適当であるが，男性については，1クラスター・モデルで充分であるとしている。Everittは，ロンドンの精神病研究所の統計学者であり，医療統計・応用統計の大家である。混合分布問題のモノグラフも出版している。しかし，このデータの分析のしかたは，疑問無しとしない。まず第一に，男性の標本平均は1.5236, 標本標準偏差は0.1734 である。つまり，男性の早発型クラスターの平均・標準偏差と，男性全体の標本平均・標本標準偏差とは，小数点以下4ケタまで一致している!さらに，ウェイトを見ると，早発型が90％を越えている。このような特徴は，早発型クラスターが，ほぼデータ全体を説明するように，推定されてしま

っていることを意味する。一方，後発型クラスターの平均は，1.8267であり，その周辺に棒グラフの突出した形状は見当たらない。これで果たして，最適の2クラスター・モデルが得られたと言えるだろうか？　このように無意味な付け足しとして，後発型クラスターが，推定されているならば，その必要性を検定によって，否定するまでもないのである。EverittはEM以外の方法で2クラスター・モデルの推定を，やり直してみるべきではなかったか？　男女のデータには，確かに共通性がある。その一方によく当てはまる2クラスター・モデルが得られたなら，より良い推定法によって，男のデータにも良く当てはまる2クラスター・モデルが得られるかもしれない。11章で検討するように，EMは万能ではないのである。

というわけで，この例は，混合分布アプローチの有効性を示すには，不満足な点が残る。しかし，敢えてそのような例を引いたのは，同じデータを別の角度から見てみたいからである。より良い推定法によって，男の方からも，女と良く似た2クラスター・モデルが得られたと仮定しよう。次に考えるべきは，これらの2クラスター・モデルが，男女共通の2クラスター・モデルからの偶然の変異に過ぎないかどうか，ということである。我々がフォーマルな取り扱い方の確立しているデータを考察するのなら，まずそのような共通モデルの説明力について考察し，それでダメだとわかってから，男女それぞれの2クラスター・モデルを推定するだろう。しかし，今は，もっと探索的に考えていることとする。さて，そのような共通モデルを考えるためには，男女のデータを一つにプールし，そこにもう一度，2クラスター・モデルを当てはめる。そして，この男女共通のモデルが男女それぞれのデータをうまく説明するかどうかの検定を行えば良い。ところが，この検定は，通常の適合度のカイ2乗検定とは，異なる。通常のカイ2乗検定を行うなら，2つのクラスターをウェイトをつけて垂直方向に足し上げ，これを推定されたモデルとして，検定を行えばよい。ところが，このやり方では2つのクラスターを，せっかく識別した意味がなくなってしまうのである。我々は2クラスター・モデルとして適合しているかどうかを調べたいのだから，1つのクラスターが非常に良く適合していても，もう1つのクラスターが大きく外れていれば，2クラスター・モデルとして適合していないと判断すべきである。ところが，非常に良く適合しているクラスターのウェイトが高い場合には，通常の検定は，もう1つのクラスターが大きく外れていても，全体として適合しているという結論を出す可能性がある。この違いは，クラスターの数が多くなれば，より深刻になるはずである。では，そのように，1つのクラスターも大きく外れていないかどうか，より精密にチェックする検定は，どう構成したら良いのか？　実は，これは，未解決の問題である。問題の設定は自然であり，男女差は人文・社会科学の到るところに現れる。男女それぞれの分布が，単純な確率分布で表わせない場合も多いだろう。このような検定が開発されれば，応用の範囲は，相当に広いに違いない。しかし，混合分布では，数理統計の通常の問題に見られるように，推定・検定・信頼区間が，密接に関連しながら発達するというスタイルで，議論が進んでいない。だから，これくらい自然でやさしく見える問題でも，解決を見ていないのである。実際のところ，Everittのアプローチの疑問点のように，推定を行う段階で，既に，克服しなければならない問題がある。そこで，我々も，本論文では推定の問題に集中する。

３．EMアルゴリズム

1章でも述べたように，今日，混合分布の推定に最もよく使われるのは，EMアルゴリズムとMCMCである。本章と次章では，クラスターの数を固定した枠組で，これらの推定法を述べ

る。EMについてはTitterington, Smith and Makov[12]が，MCMCについてはGilks, Richardson and Spiegelhalter [4]が，簡潔な展望を与えている。そこで，我々も，これらの本の記述に沿いつつ，より平易で，二つの手法の対比が明瞭な解説を試みる。尚，混合分布の中の一つ一つの小さな確率分布は，クラスターともコンポーネントとも呼ばれる。「クラスター」と言えば「分類的」な，「コンポーネント」と言えば「構成要素的」なニュアンスが伴うが，このような違いには余りこだわらず，適宜，同じ意味で使っていく。また，単純で紙数を費やす計算が現れるときは，細かい部分に立ち入らない。

観測値がn個あるデータに，kコンポーネントの混合分布を当てはめる問題を考えよう。この分布全体は

$$g(x \mid \psi) = \sum_{j=1}^{k} \pi_j f(x \mid \theta_j)$$

と表せる。それぞれのコンポーネントは，同じ型で異なるパラメターの確率密度関数fを持っている。j番目のコンポーネントのウェイトはπ_j，パラメター・ベクトルはθ_jである。全てのコンポーネントのウェイトとパラメターを，まとめて表現したいときは，文字ψを用いる。確率変数xについては，もし読者が，多次元分布の期待値の計算になじみがなければ，1次元と思って構わない。しかし，本章では，xが何次元でも通用するように，記述を調整してある。そこで，対数尤度関数は，

$$L(\psi) = \sum_{i=1}^{n} \log\left[\sum_{j=1}^{k} \pi_j f\left(x_i \mid \theta_j\right)\right]$$

ここで，現実のデータとしては得られない，仮説的な変数$z_i \quad i = 1, \cdots\cdots, n$を設定する。各$z_i$はk次元の縦ベクトルである。$z_i$の各要素は，$z_{ij} \quad j = 1, \cdots\cdots, k$で，添え時は通常の場合と逆順である。$z_i$はいわゆるindicator vectorであり，x_iが第jコンポーネントから生み出された場合に，$z_{ij} = 1$。それ以外のz_iの要素は全て0である。このようなz_iがすべてのx_iに対応してわかっているのであれば，

$$\{y_i, i = 1, \cdots\cdots, n\} = \{(x_i, z_i), i = 1, \cdots\cdots, n\}$$

と書けるだろう。ペア(x_i, z_i)の単なる代理として，y_iと書く。$\underset{\sim}{y} = (y_1, \cdots\cdots, y_n)$に基ずく尤度関数は，

$$\lambda(y_1, \cdots\cdots, y_n \mid \psi) = \prod_{i=1}^{n} \prod_{j=1}^{k} \pi_j^{z_{ij}} f_j\left(x_i \mid \theta_j\right)^{z_{ij}}$$

対数を取って，

$$\Lambda\left(\underset{\sim}{y} \mid \psi\right) = \sum_{i=1}^{n} z_i^T A(\pi) + \sum_{i=1}^{n} z_i^T B_i(\theta)$$

ここに，$A(\pi)$ のj番目の要素は $\log \pi_j$ ，$B_i(\theta)$ のj番目の要素は $\log f_j\left(x_i \mid \theta_j\right)$ である。

Eステップ（期待値の計算）：$E\left[\Lambda\left(\underset{\sim}{y} \mid \psi\right) \mid \underset{\sim}{x}, \psi^{(m)}\right] = U\left(\psi, \psi^{(m)}\right)$

なる期待値の関数形を求めよ。

我々は，現在，イタレーションの第m+1段階を処理中であるとしよう。すると，$\psi^{(m)}$ は前段階で得られた全てのパラメターということになる。ψ に含まれる π や θ を未知のパラメターにしたまま，Λ の条件付き期待値を求めよ，というのが，ここでの問題の意味である。一見すると，そんなことは絶望的に見える。ところが，以下のように考えれば，この関数形は求まるのである。すでに求めた $\Lambda\left(\underset{\sim}{y} \mid \psi\right)$ から，

$$U\left(\psi, \psi^{(m)}\right) = \sum_{i=1}^{n} w_i\left(\psi^{(m)}\right)^T A(\pi) + \sum_{i=1}^{n} w_i\left(\psi^{(m)}\right)^T B_i(\theta)$$

ここに，$w_i\left(\psi^{(m)}\right) = E\left(z_i \mid x_i, \psi^{(m)}\right)$
この左辺は z_i に対応するk次元ベクトルで，そのj番目の要素は，

$$\begin{aligned} & w_{ij}\left(\psi^{(m)}\right) \\ & = \left[w_i\left(\psi^{(m)}\right)\right]_j \\ & = \pi_j^{(m)} \frac{f_j\left(x_i \mid \theta_j^{(m)}\right)}{g\left(x_i \mid \psi^{(m)}\right)} \end{aligned}$$

最後のステップは，

$\pi_j^{(m)}$ ＝第m段階における第jコンポーネントのウェイト

$f_j\left(x_i \mid \theta_j^{(m)}\right)$ ＝第m段階のパラメターを前提とした，第jコンポーネントの x_i における値

$g\left(x_i \mid \psi^{(m)}\right)$ ＝第m段階の全パラメターを前提とした，kコンポーネント・モデルの x_i における値

という式の意味による。このように計算を進めて行くためには，もちろん，第1段階のため

に，$\psi^{(0)}$ を適当に与えておかなくてはならない。しかし，それさえ与えられれば，z_i に関して何も確実な情報が無いにもかかわらず，その条件付き期待値は求まるというのが，トリックのポイントである。

Mステップ（最大化）: $U\left(\psi,\psi^{(m)}\right)$ を最大化する ψ を求めよ。

このように，各段階ごとに，EステップとMステップを繰り返しながら，EMアルゴリズムは進んで行く。即ちExpectation Maximization Algorithmである。このアルゴリズムの段階を進むにつれて，尤度は単調増大することが，証明されている。しかし，大域的最大値は保証されない。むしろ，この尤度関数は，たくさんの局所的最大値を持っている。なぜならば，モデルに取りこまれた，たくさんのパラメターが，全て変数として扱われる，多変数関数だからである。初期値 $\psi^{(0)}$ の選択に関しては，幾つかの提案がある。しかし，アルゴリズムによって，最終的に選択された ψ が，初期値 $\psi^{(0)}$ によって異なっていたという，報告は多い。このような問題点から，EMは完璧には程遠いと考えられている。

４．MCMC

MCMC（マルコフ連鎖モンテカルロ法，Markov Chain Monte Carlo）は，混合分布へのベイズ的アプローチの代表的なものである。しかしながら，MCMCは，基本的な考えを，EMから借りているようなところがあり，その共通点とMCMC独自の点とを，明らかにしなければならない。そこで，前章で使用した文字も別の意味で再定義するので，注意して読んで欲しい。混合分布全体は

$$g(x)=\sum_{i=1}^{k} p_i f(x\mid\theta_i)$$

と書く。ウェイトは p_i，iとj は前章と逆の意味である。各コンポーネントは非常に広く，exponential familyとしよう。

$$f(x\mid\theta)=c(x)e^{\theta\cdot x-Q(\theta)}$$

と書けば，共役な事前分布が考えられ，

$$\pi(\theta\mid\nu,\lambda)\propto e^{\theta\cdot\nu-\lambda Q(\theta)}$$

ここに ν は定数のベクトルで，θ と同じ長さを持ち，λ はスカラーの定数である。ウェイト $p_1,\cdots\cdots,p_k$ の共役な事前分布はディリクレ分布 $D(r_1,\cdots\cdots,r_k)$ で与えられるとしよう。すると，その密度関数は

$$\tau(p_1,\cdots\cdots,p_k) \propto p_1^{r_1-1}\cdots\cdots p_k^{r_k-1}$$

これらによって，事後分布は，

$$h_1(p_1,\cdots\cdots,p_k,\theta_1,\cdots\cdots,\theta_k \mid x_1,\cdots\cdots,x_n) \propto$$

$$\tau(p_1,\cdots\cdots,p_k)\prod_{i=1}^{k}\pi(\theta_i \mid \nu_i,\lambda_i)\prod_{j=1}^{n}\left(\sum_{i=1}^{k} p_i f(x_j \mid \theta_i)\right)$$

ここで，z変数を導入するが，前章とは異なる。ベクトル$(z_1,\cdots\cdots,z_n)$において，z_jは，x_jが生成されたコンポーネントの番号を表しているとする。もし，このようなz_jの値が全てわかっているのなら，h_1の式から非常に多くの項を省くことができる。

$$h_2(p_1,\cdots\cdots,p_k,\theta_1,\cdots\cdots,\theta_k \mid x_1,\cdots\cdots,x_n,z_1,\cdots\cdots,z_n) \propto$$

$$p_1^{r_1+n_1-1}\cdots\cdots p_k^{r_k+n_k-1}\pi(\theta_1 \mid \nu_1+n_1\bar{x}_1,\lambda_1+n_1)\cdots\cdots\pi(\theta_k \mid \nu_k+n_k\bar{x}_k,\lambda_k+n_k)$$

ここに　　$n_i=\sum_j I(z_j=i)$　　$n_i\bar{x}_i=\sum_{j:z_j=i} x_j$

Iは（　）内の条件が満たされているとき1，そうでないとき0を取るindicator functionである。以上，準備したことから，次のようにイタレーションを構成する。

ステップ0：$z_1,\cdots\cdots,z_n$に初期値を与え，これに基づいてn_iと$\bar{x}_i$ $(i=1,\cdots\cdots,k)$を計算する。

ステップ1：パラメター・ベクトルθ_iとウェイトp_i $(i=1,\cdots\cdots,k)$を以下の分布形に基づいて生成する。

$$\theta_i \sim \pi(\theta_i \mid \nu_i+n_i\bar{x}_i,\lambda_i+n_i) \qquad i=1,\cdots\cdots,k$$

$$(p_1,\cdots\cdots,p_k) \sim D(r_1+n_1,\cdots\cdots,r_k+n_k)$$

ステップ2：z_j $(j=1,\cdots\cdots,n)$を以下の分布形に基づいて生成する。

$$h_3(z_j \mid x_j,p_1,\cdots\cdots,p_k,\theta_1,\cdots\cdots,\theta_k)=\sum_{i=1}^{k} p_{ij} I(z_j=i) \qquad j=1,\cdots\cdots,n$$

$$\text{ここに}\quad p_{ij}=\frac{p_i f(x_j \mid \theta_i)}{\sum_{t=1}^{k} p_t f(x_j \mid \theta_t)} \qquad i=1,\cdots\cdots,k$$

ステップ3：n_i と $\bar{x}_i \quad i=1,\cdots\cdots,k$ を更新する。

パラメターと z_j の生成に関しては，ベイジアンの枠組にふさわしい方式が開発されており，Gibbs sampling という（Gilks, Richardson，Spiegelhalter[4]参照）。このイタレーションを続けていけば，以下の近似が充分よくあてはまる，イタレーション回数に達する。

$$E[T(\theta,p)] \approx \frac{1}{M}\sum_{i=1}^{M} T\left(\theta^{(i)},p^{(i)}\right)$$

ここで $i=1,\cdots\cdots,M$ は，イタレーション回数が，今述べた境界をM回越えていることを意味する。$\theta^{(i)}$，$p^{(i)}$ は，この意味での第i回イタレーションで生成されたθとpである。Tはθとpを変数とする任意の関数である。

ベイジアンの視点が一貫してはいるが，MCMCは内在的な問題から免れているわけではない。特に深刻なのは「ラベル転換問題」である。これまで仮定した通り，各コンポーネントは同じ型の確率分布に属している。この場合，事後分布は，コンポーネント番号（ラベル）を入れ替えても不変である。イタレーションの過程で，コンポーネント番号がふり替わってしまうことは，実際にあり得るのであり，しかも，このことはprocedure の収束を妨げるほど重大な問題である。我々はこの問題に深く立ち入らないが，まだ満足な解決を見ていないということは，言っておこう。

MCMCはEMよりも良いアルゴリズムなのだろうか？　文献を調べてみると，この重要な問題に関して，わずかなことしかわかっていないのに驚く。現状を語るには，「競争そのものが，うまく formulate されていない。」としか言いようがない。行きづまりの原因として，少なくとも三つのことが，言えるだろう。第一に，混合分布の問題では，パラメターの数が非常に多く，どんなアプローチを取っても，計算の過程は長い。このような問題では，通常の理論統計学的基準は，直観的なアピールが乏しい。かと言って，それに代わる，混合分布にふさわしい，理論的基準が提案されたわけでもない。第二に，MCMCはベイジアン・アプローチの理論統計学者に好まれるだろうが，利用するユーザーにとっては，不便な点がある。EMでおかしな答えが出てきたときに，ユーザーは，別な初期値を試してみるだろう。しかし，ラベル転換問題の場合には，ユーザーが，その場に応じて処理できるような，解決策がない。第三に，シミュレーションに基づく比較研究も，積極的に行われたわけではなかった。統計の問題で，このような研究が手薄になるのは奇妙なことである。しかし，これには次のような事情があろう。混合分布は非常に大きなバラエティーを許容するので，どのくらいの難易度を，この分野のゴールとするのか，定め難い。そこで，異なった方法をテストするための，代表的な実例を選ぶのも困難なのである。

5．Scott と Szewczyk による新しいアプローチ

前二章で説明した，EMアルゴリズムとMCMCは，クラスター数の固定されたモデルの推定を，主要な目標としている。クラスター数を推定するには，これらのアプローチは，何等かの情報量基準に頼らざるを得ない。まず，あり得ると思われる，クラスター数の範囲を定める。

次に，その一つ一つについて，クラスター数を固定した推定を行う。この推定値に基づいてICを計算し，これが最適（普通，最小）となるクラスター数を選択する。従って，モデル・パラメターの推定と，クラスター数の選択は，基本的に別々の作業ということになる。Titterington, Smith, and Makov（1985）[12]とMcLachlan and Peel（2000）[8]に載せられている文献を比較してみれば，クラスター数を固定したモデルの推定に関しては，非常にたくさんの研究があるが，クラスター数の決定に関しては，それほど大きな進歩はなかったことがわかる。ICを用いたアプローチの改良が，その主なものである。さらに，前二章で述べた，EMとMCMCの難点に加えて，もう一つの問題を心に留めておかなければならない。それは，計算の実行速度である。最近では，高次元（例えば100次元）で大規模（例えば観測値数100,000）なデータ処理の需要が急増しており，EMの処理速度の遅さが，新たな懸念を生んでいる。クラスター数選択の実際は，単純に，一般的に述べることができる。procedureが，あらゆるクラスター数を通して，ある程度良いフィットを維持し，正しいクラスター数において，取り分け良いフィットを達成するなら，ICは，この正しいクラスター数を選ぶだろう。あらゆるクラスター数において，できる限りフィットが高くなるように設計されたアルゴリズムでは，クラスター数選択のときに，過剰な候補をかかえることになる。（このことの意味は，本論文の11章で明らかとなる。）これらの理由，即ち，既存の方法に残存する問題，大規模多次元データへの対応，クラスター数選択の基本的に簡明な構成は，我々をして，既存アプローチの彫琢よりもむしろ，単純化に向かわしめる。このような方向に最初に乗り出したのが，ScottとSzewczykの論文[11]である。

ScottとSzewczyk（SSと略記）は，彼らのprocedureを2001年に発表したが，その真価は，まだ充分に認められていない。これには，二つの理由が考えられる。第一に，既存の方法と比べて，彼らのアプローチが単純なように見えることである。そこで，混合分布問題（特にクラスター数の決定）を特別な難問と思っている人達に，見過ごされる傾向があった。しかし，実際は，本論文を読み進めばわかる通り，彼らのアルゴリズムと我々によるその拡張のパフォーマンスは，良好である。「これまでの研究者の期待を，はるかに超える」と言っても良いくらいである。SSの論文の第二の問題としては，多くの新しいアイディアが試みられているにもかかわらず，それらは必ずしも良く説明されていないということがある。彼らの新しい概念装置が，混合分布の本質を深くとらえているということを，納得するには，SSの原文をよりていねいに説明し，たくさんのシミュレーションを行わなければならない。当然のことながら，これら，克服すべき点を考慮に入れて，本論文の議論は組み立てられることになる。しかし，それをどうするかを述べる前に，我々は，SSのprocedureを簡単に紹介しておこう。ただし，Phase2は彼らのオリジナルそのままではなく，本論文の議論の構成にふさわしいように，変更が加えてある。こうしたのは，彼らはPhase2において，計算の能率を優先して，省略計算を取りこみ過ぎているからである。結果として，彼らのprocedureでは，Phase2の新概念がPhase3及び4の新概念と比較できないようになっている。我々のPhase2はより原始的（つまり，SSの元の意図に忠実）ではあるが，それ以後のphaseとの比較が容易になっている。本章ではまた，procedureの解説を要点だけに留める。新概念の意味を充分に納得するためには，例が必要である。しかし，2次元の実例は極めて咀嚼しやすいことがわかった。そこで，1次元procedureの詳細は，6-8章で2次元への拡張を考察する際に，もう一度掘り下げる。SSのprocedureは4部から成っている。

表２．SSの1次元procedure

1	カーネル推定	
	推定法	相似測度
2	MM	1 vs. 1
3	MM	*n* vs. *n* -1
4	L2E 最適化	*n* vs. *n* -1

Phase1はカーネル推定である。我々は，どのカーネルも同じ大きさの包摂バンド（bandwidth, 「バンド幅」と訳されることがあるが，日本語として座わりが悪いので，「包摂バンド」と訳す）を持った，標準的なカーネル推定を行う。包摂バンドの選択法はクロス・バリデーションである。

$$\hat{f}(x;h)=\frac{1}{N}\cdot\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{h}\cdot\exp\left\{-\frac{1}{2}\left(\frac{x-X_i}{h}\right)^2\right\}$$

ここにhは包摂バンド，X_iはi番目の観測値，Nは観測値数である。このようにカーネルによって密度関数を推定した結果は，Nコンポーネントの混合分布とみなし得る。我々は以下のようにコンポーネント数を削減して，Nコンポーネント・モデルを最適なコンポーネント数のモデルに近づけて行く。削減プロセスは3つのphaseに分かれる。その最初のPhase2では，最も計算効率の高い方法が用いられる。まず，コンポーネント同士の相似性を相似測度によって評価する。

$$\frac{\int_{-\infty}^{\infty}f_1(x)f_2(x)dx}{\sqrt{\int_{-\infty}^{\infty}f_1^2(x)dx}\sqrt{\int_{-\infty}^{\infty}f_2^2(x)dx}}$$

この測度は，2つのコンポーネント（どちらもpdfである）の位置と形を両方取り入れた「相似性」を測っていることを，注意しておこう。次に，相似測度の値が最高のペアを選び，これを1つのコンポーネントに融合する。融合法としては，積率法（Method of Moments, MMと略）を用いる。第一のコンポーネントは$N(\mu_1,\sigma_1)$でウェイトはw_1，第二のコンポーネントは$N(\mu_2,\sigma_2)$でウェイトはw_2としよう。すると，新しいコンポーネントのパラメターとウェイトは

$$\mu_{new} = \frac{w_1}{w_1 + w_2}\mu_1 + \frac{w_2}{w_1 + w_2}\mu_2$$

$$\sigma^2_{new} = \frac{w_1}{w_1 + w_2}\sigma_1^2 + \frac{w_2}{w_1 + w_2}\sigma_2^2 + \left(\frac{w_1}{w_1 + w_2}\right)\left(\frac{w_2}{w_1 + w_2}\right)(\mu_1 - \mu_2)^2$$

$$w_{new} = w_1 + w_2$$

SSは，この式が何に基づいているのか，明らかにしていない。しかし，我々は，7章において詳しく論ずる。ここで*N*段階は終了し，*N*-1段階に入る。そこでは，*N*-1コンポーネントを*N*-2コンポーネントに削減する，同じ作業がくり返される。このようにして，クラスター数nを下って行き，Phase2と3の境界のnに達する。この境界とPhase3から4への，もう一つの境界は，アド・ホック的に選ばれている。しかし，SSの例と本論文のシミュレーションから見る限り，これらは適切に選ばれている。Phase3では，融合法として，MMを残すが，相似測度を変更する。今，nクラスター・モデルを処理中で，n-1クラスター・モデルに減らすところだ，としよう。簡単のために，隣り合ったコンポーネント（隣り合っているかどうかは，例えば，各コンポーネントの平均の位置で決める）のみを考え，以下のやり方で，融合するのに最も良いペアを選ぶ。まず我々は，コンポーネント番号1と2を1つに融合する。nクラスター・モデルの内，番号1,2以外の全てのコンポーネントを残し，これに融合したコンポーネントを加えて，n-1クラスター・モデルの候補が一つできあがる。次に，nクラスター・モデルと，この候補モデルの間の相似測度を計算する。ここで用いる相似測度とは，nクラスター vs. n-1クラスター相似測度である。個々のクラスター同士の相似測度に似せて，新しい相似測度は

$$\frac{\int_{-\infty}^{\infty}\lambda(x)\gamma(x)dx}{\sqrt{\int_{-\infty}^{\infty}\lambda^2(x)dx}\sqrt{\int_{-\infty}^{\infty}\gamma^2(x)dx}}$$

と定義される。ここにλはnクラスター・モデル全体の，γはn-1クラスター・モデル全体のpdfである。次に，もう一度nクラスター・モデルに戻り，そのコンポーネント番号2と3を1つに融合する。コンポーネント番号2,3以外の全てのコンポーネントを残し，もう一つの候補モデルを作る。それから，相似測度を計算する。このようにして進んで行き，全てのペアについてやり終えた後，nクラスター・モデルとの間の相似測度が最高となる候補モデルを，n-1のクラスの最適のモデルとして選択する。

詳しい説明抜きで，ザッとPhase3を眺めただけでも，Phase2と3の違いは何なのか，1 vs. 1相似測度と，n vs. n-1相似測度の違いは何なのか，ということに戸惑いを覚える。ところが，本研究によって，この違いは，極めて重要であることが，わかった。そこで，実際のprocedureの過程の中で，この違いをどうやって確かめるか，それをどう説明し，理論的背景をどう確立するかという問題は，本論文の焦点の一つとなる。

適当な境界を越した後，Phase4が始まる。ここでも我々は，クラスター数nを降下していくプロセスを続けるが，もう最適のクラスターに近いと考えられるので，計算時間を費やす方法を用いてもさしつかえない。n vs. n-1相似測度は，そのままに残すが，融合法はL2E最適化に置き換える。L2EはDavid W. Scottによって新しく提案された，パラメター推定の原理である。今，考えている混合分布の枠組では，outlierに対する頑健性を，この方法のメリットと，彼は見なしている。積分

$$\int\left\{\hat{f}(x;\theta)-f(x)\right\}^2 dx$$

において，$\hat{f}$ は推定すべき密度関数，fは真の密度関数，θは推定の目標たるパラメター・ベクトルである。L2Eはこの積分を最小化する $\hat{f}$ を求めることでfを推定しようとする。
この式は真の密度関数を含んでいるので，直接には操作可能でない。しかし，

$$\int\hat{f}(x;\theta)^2 dx-\frac{2}{N}\sum_{i=1}^{N}\hat{f}(X_i;\theta)$$

によって近似できることが，わかっている。さらに，正規混合分布の場合，一項目は，積分記号の入らない閉じた形に変形できる。今，紹介している論文とは別に，Scottは，L2Eのためのもう一編の論文[10]を著わしているので，その理論的な性質については，そちらを参照して欲しい。Phase4では，各クラスター数において選ばれたモデルから，情報量基準を計算する。これに基づいて，最適なクラスター数を選び，そのクラスター数で選ばれていたモデルが，最適なモデルである。

[11]において述べられている，その他の結果を，簡単に要約しておこう。彼らは，このprocedureを，コンピュータで実行する上での，細かい問題点について論じている。ただし，1次元のみである。経済学および天文学のデータへの応用を試みている。仮説例から発生させたデータ・セットについては，三つの例を挙げている。ただし，これらは，うまくいった例だけである。一つの例から，たくさんのデータ・セットを発生させて行う，繰り返し実験は，やっていない。高次元データについても，試みたようであるが，方法の詳細は示されていない。彼らのprocedureが，どれくらいうまくいったかについての，定量的な評価は，充分に行われていない。繰り返しシミュレーションによるデータ・セット中，いくつに対して，正しいクラスター数が得られたか，成功率を示す必要があろう。このことは重要である。混合分布のような分野では，全ての方法を評価する基準が確立していない。個々の方法の得失を知るために，多様なシミュレーションを試みなければならない。

SSのprocedureをカーネル降下法（Kernel Reduction, KR）と呼ぼう。彼らの達成したことを踏まえて，本論文では，二つの新しい課題を設定する。第一は，彼らが充分に説明できなかった，新概念の意味を明確にすることである。第二は，1次元正規混合分布のためのprocedureを，2次元正規混合分布に拡張することである。すでに述べたように，2次元の例はわかりやすいので，これからの叙述は，第二の課題を答える順通りとし，第一の課題は，その中の適当な所で答えることとする。尚，Phase3は，データ・サイズもクラスター数も少ない場合には，省略

できることがわかった。そこで，本論文ではPhase3を飛ばし，Phase4をPhase3と呼ぶことにする。この3 phase procedureはKRの基本的なアイディアを全て含んでいる。

６．２次元のためのPhase1（カーネル推定）

SSのオリジナル・アプローチでは，全てのカーネルに共通の包摂バンドを仮定して，カーネル推定が行われる。通常，これを2次元に拡張するときは，3つのパラメター σ_x，σ_y，ρ を推定する。（各カーネルの(μ_x，μ_y)は観測値によって置き換えられる。）しかし，我々は，これよりも単純なアプローチを試みる。即ち，ρ=0と置いて，$\sigma=\sigma_x=\sigma_y$ のみを推定するのである。その理由を述べよう。真のモデルが，数個のクラスターから成っていると仮定する。それらの全ては，一つを除いて，右肩上がりの等高線（正の相関）を持っているとしよう。例外のクラスターは左肩上がりの等高線（負の相関）を持っている。すると，カーネルもまた，右肩上がりと推定されるに違いない。このことは，ほとんどのクラスターの降下過程にとって，有利な背景となる。しかし，一つだけの例外のクラスターにとっては，問題を生じるだろう。降下過程を進むに従って，クラスターの本来の形が形成されてくる。これは，例外のクラスターの場合，もちろん，左肩上がりである。しかし，形成されつつあるクラスターと，残りのカーネルの間の相似測度の値は小さい。何故ならば，形が違うからである。この困難は，高次元では，更に大きくなる。高次元では，より多くの直交性が存在するのだから，カーネルが，統合されるべきクラスターに対して，直交に近く向いている可能性も大きくなる。カーネル推定のたいていの文献で，ρは推定されているが，その意味について，特に筋の通った解釈がなされているわけではない。どんな場合にρをモデルに含み，どんな場合に含めないかということは，もっと注意を向けられて良い問題である。

以上述べてきた推定法は，最も標準的なカーネル推定を，若干修正したものにすぎない。だから，カーネルについて良く知っている読者には，詳しく展開する必要は，あるまい。しかし，我々の降下procedureが，どんなところから始まるかを，明らかにするために，数学的な形を示そう。

今，N個の観測値があるとしよう。カーネルによって推定された確率分布は，

$$\hat{f}(x;h)=\frac{1}{N}\sum_{i=1}^{N}K_h(x-X_i)$$

で与えられる。$K_h(x-X_i)$は，観測値X_iの「上に乗っている」カーネル，即ち，平均X_iの2次元正規分布である。xとyの両方向に等しい分散は，hで表わされている。ベクトルxとX_iを

$$x=\begin{pmatrix}x_1\\x_2\end{pmatrix}\qquad\qquad X_i=\begin{pmatrix}X_{i1}\\X_{i2}\end{pmatrix}$$

と書こう。我々はσ_x，σ_y，ρ,に関する単純化の仮定を持っているから，$K_h(x-X_i)$,は

$$\frac{1}{2\pi}\cdot\frac{1}{h^2}\cdot\exp\left(-\frac{1}{2}\left\{\left(\frac{x_1-X_{i1}}{h}\right)^2+\left(\frac{x_2-X_{i2}}{h}\right)^2\right\}\right)$$

と書ける。

$$\therefore \qquad \hat{f}(x;h) = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{1}{2\pi} \cdot \frac{1}{h^2} \cdot \exp\left(-\frac{1}{2} \left\{ \left(\frac{x_1 - X_{i1}}{h} \right)^2 + \left(\frac{x_2 - X_{i2}}{h} \right)^2 \right\} \right)$$

包摂バンドの決定には，最小二乗クロス・バリデーション（Least Squares Cross-Validation, LSCV）が最も良く用いられる。LSCVは二乗誤差積分平均（Mean Integral Squared Error, MISE）の最小化を目指している。

$$MISE(h) = E \int \left(\hat{f}(x;h) - f(x) \right)^2 dx$$

$$= E \int \hat{f}(x;h)^2 dx - 2E \int \hat{f}(x;h) f(x) dx + \int f(x)^2 dx$$

ここに $f(x)$ は真の密度関数である。最後の項は最小化すべきパラメターを含んでいないから，最初の二項のみを問題にする。以下の式が，最初の二項の不偏推定量を与えることがわかっている。

$$LSCV(h) = \int \hat{f}(x;h)^2 dx - \frac{2}{N} \sum_{i=1}^{N} \hat{f}_{-i}(X_i;h)$$

ここに $\hat{f}_{-i}(x;h) = \frac{1}{N-1} \sum_{j \neq i}^{N} K_h(x - X_j)$

LSCVの一項目は，2次元正規分布の積の積分を含んでいるが，これは以下の公式（Wand and Jones [14] 参照）によって簡単に求まる。

$$\int \phi_{\Sigma}(x - \mu) \phi_{\Sigma'}(x - \mu') dx = \phi_{\Sigma + \Sigma'}(\mu - \mu') \qquad (1)$$

平均μ，分散共分散行列Σのd次元正規分布の記号が $\phi_{\Sigma}(x - \mu)$ である。これより，LSCVの二項の具体的な形は，

$$\int \hat{f}(x;h)^2 dx = \frac{1}{N^2} \left\{ N \cdot \left(\frac{1}{2\sqrt{\pi} h} \right)^2 + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \varphi(i,j,1) \varphi(i,j,2) \right\}$$

ここに $\varphi(i,j,k) = \frac{1}{2\sqrt{\pi} h} e^{-\frac{1}{4} \left(\frac{X_{ik} - X_{jk}}{h} \right)^2}$

$$-\frac{2}{N}\sum_{i=1}^{N}\hat{f}_{-i}(X_i;h)$$

$$=-\frac{2}{N}\cdot\frac{1}{N-1}\cdot\frac{1}{h^2}\cdot 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{1}{2\pi}\exp\left(-\frac{1}{2}\left\{\left(\frac{X_{i1}-X_{j1}}{h}\right)^2+\left(\frac{X_{i2}-X_{j2}}{h}\right)^2\right\}\right)$$

７．２次元のための Phase2（MM+MEASURE1）

Phase2では，積率法（Method of Moments, MM）と1 vs. 1クラスター相似測度によって，クラスター数を降下させて行く。1 vs. 1クラスター相似測度をMEASURE1，n vs. n-1 クラスター相似測度をMEASURE2と呼ぼう。1次元では，MEASURE1の形はSSによって，

$$sim1(f_1,f_2)=\frac{\int_{-\infty}^{\infty}f_1(x)f_2(x)dx}{\sqrt{\int_{-\infty}^{\infty}f_1^2(x)dx}\sqrt{\int_{-\infty}^{\infty}f_2^2(x)dx}}$$

と与えられている。この式が $0\le sim1\le 1$ の値をとることはコーシー＝シュワルツ不等式によってわかる。この測度の具体的な形を求めるには，6章の（1）式を使うことができる。そこで，

$$=\frac{(2\sigma_1\sigma_2)^{1/2}}{(\sigma_1^2+\sigma_2^2)^{1/2}}\exp\left[\frac{-(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right]$$

ここに，$f_1\sim N(\mu_1,\sigma_1^2)$　$f_2\sim N(\mu_2,\sigma_2^2)$　$sim1$は，ほとんど形を変えずに直ちに2次元に拡張される。

$$sim2(f_1,f_2)=\frac{\iint f_1(x,y)f_2(x,y)dydx}{\sqrt{\iint f_1^2(x,y)dydx}\sqrt{\iint f_2^2(x,y)dydx}}$$

ここに，全ての積分記号は－∞から∞に及ぶ定積分である。この数量も $0\le sim2\le 1$ を満足する。tを実数としよう。

$$\{f_1(x,y)+tf_2(x,y)\}^2\ge 0$$

$$\therefore\quad f_1(x,y)^2+2tf_1(x,y)f_2(x,y)+t^2f_2(x,y)^2\ge 0$$

$$\therefore\quad \iint f_1^2(x,y)dydx+2t\iint f_1(x,y)f_2(x,y)dydx+t^2\iint f_2^2(x,y)dydx\ge 0$$

最後の式の左辺は，tに関して放物線である。

$$\therefore \quad D = \left\{\iint f_1(x,y) f_2(x,y)\,dydx\right\}^2 - \left\{\iint f_1^2(x,y)\,dydx\right\}\left\{\iint f_2^2(x,y)\,dydx\right\} \le 0$$

$$\left\{\iint f_1(x,y) f_2(x,y)\,dydx\right\}^2 \le \left\{\iint f_1^2(x,y)\,dydx\right\}\left\{\iint f_2^2(x,y)\,dydx\right\}$$

よって，

$$0 \le \frac{\iint f_1(x,y) f_2(x,y)\,dydx}{\sqrt{\iint f_1^2(x,y)\,dydx}\sqrt{\iint f_2^2(x,y)\,dydx}} \le 1$$

*sim*2 の具体的な形は，再び（1）によって求まる。

$$sim2(f_1,f_2) = \frac{\left(|2\Sigma_1|\cdot|2\Sigma_2|\right)^{1/4}}{|\Sigma_1+\Sigma_2|^{1/2}} \exp\left[-\frac{1}{2}(\mu_1-\mu_2)^T(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)\right]$$

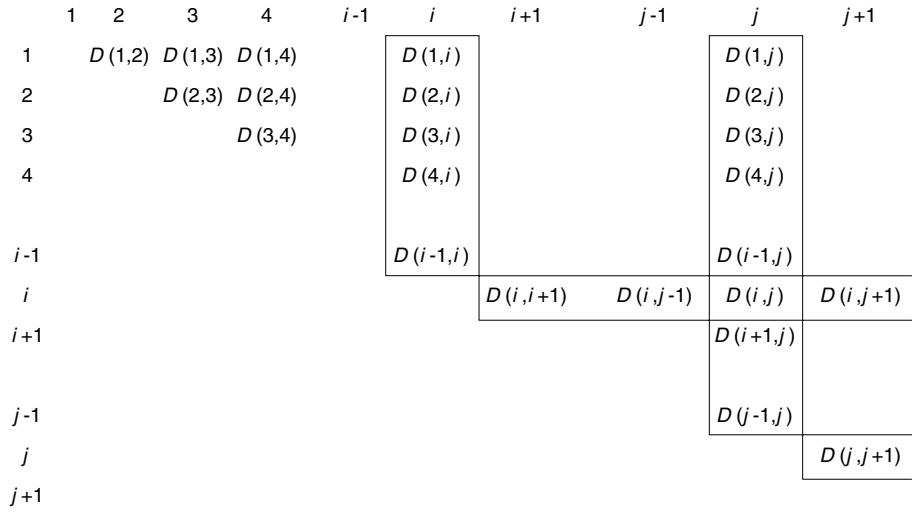
ここに $f_1 \sim N(\mu_1,\Sigma_1)$　$f_2 \sim N(\mu_2,\Sigma_2)$

相似測度ができあがった後，我々は，どのコンポーネントを取り除くかの決定に進む。既に5章で触れたように，SSの方法は，この点に関してやや混乱している。彼らは，*sim*1そのものを計算せず，もっと簡単に計算できる量で置き換えて構わない，と言うのである。しかし，このやり方は，新概念の比較を不可能にするだけでなく，混合分布の応用の一般的な目的に照らして，問題がある。本論文では，紙幅の関係から，この点について詳論しないが，関心ある読者はKaneda[6] のp.24～26を参照されたい。

省略計算に頼らず，相似測度をそのまま計算するなら，計算量の問題に直面する。しかし，これは，効率良い簿記を工夫することによって，かなり良く解決できることがわかった。このアイディアは，数学的トリックを含まず，単なる簿記に過ぎない。しかし，混合分布ではどんなアプローチを取るにしても，計算時間が問題になる。我々の単純な工夫も，アルゴリズムの実行には不可欠と思われるので，ここに紹介する。

観測値の数をNとしよう。Phase2は，カーネル推定によって得られたNコンポーネント・モデルから出発する。コンポーネントの可能な全てのペアについて，相似測度が計算される。この$\binom{N}{2}$個の相似測度は，次の図の三角行列に並べられる。

図３．三角行列の左上部



この全体をサーチした結果，i番目のコンポーネントとj番目のコンポーネントの間の相似測度が，最大であったとしよう。第iコンポーネントを相手として含む，全てのペアと，第jコンポーネントを相手として含む，全てのペアの相似測度が四角で囲まれている。三角行列中で，これらの要素だけが，第iコンポーネントと第jコンポーネントの融合によって影響を受ける。一般性を失うことなく，$i<j$と仮定し，以下の規則を設ける。j行とj列を消去し，i行とi列は，新しく融合されたコンポーネントと，残っているコンポーネントの間の相似測度で置き換える。行列のサイズの調整のために，j列よりも右側の列は，全て1列左に移動し，j行よりも下側の行は，全て1行上に移動する。これで，N-1段階における，最大の相似測度をサーチする，準備ができたことになる。つまり，N-1段階では$\frac{(N-1)(N-2)}{2}$個の全相似測度を計算する必要はなく，N-2個で充分である。我々の実例では，Phase3は30コンポーネント段階から始まるが，観測値数400と2000の場合，Phase2の所要時間は，それぞれ3分と13分に過ぎない。したがって，この節約は，多くの応用において充分な力を発揮すると考えられる。

融合法について考察しよう。1次元では，SSは，新しく融合されたコンポーネントのパラメターとウェイトは

$$\mu_{new} = \frac{w_1}{w_1 + w_2}\mu_1 + \frac{w_2}{w_1 + w_2}\mu_2$$

$$\sigma^2_{new} = \frac{w_1}{w_1 + w_2}\sigma_1^2 + \frac{w_2}{w_1 + w_2}\sigma_2^2 + \left(\frac{w_1}{w_1 + w_2}\right)\left(\frac{w_2}{w_1 + w_2}\right)(\mu_1 - \mu_2)^2$$

$$w_{new} = w_1 + w_2$$

であるべきだと主張する。彼らは，この公式が何から出て来るのか，語っていない。しかし，次の事実（ホーエル[5], p.113参照）を知れば，その根拠が明らかになる。

大きさn_1とn_2の二群の観測値があるとしよう。各々の平均と標準偏差を，$\bar{x}_1, \bar{x}_2$およびs_1, s_2とする。この二群を融合した一つのデータの平均と標準偏差を求めよ。答えは

$$\bar{x} = \frac{n_1}{n_1 + n_2}\bar{x}_1 + \frac{n_2}{n_1 + n_2}\bar{x}_2$$

$$s^2 = \frac{n_1}{n_1 + n_2}s_1^2 + \frac{n_2}{n_1 + n_2}s_2^2 + \left(\frac{n_1}{n_1 + n_2}\right)\left(\frac{n_2}{n_1 + n_2}\right)(\bar{x}_1 - \bar{x}_2)^2$$

我々が，混合分布の二つのコンポーネントから，合わせてサイズnのサンプルを生成するなら，$\frac{w_1}{w_1 + w_2}n$は第一のコンポーネントから，$\frac{w_2}{w_1 + w_2}n$は第二のコンポーネントから来ることになる。これらを上の公式に代入すると，

$$\bar{x} = \frac{w_1}{w_1 + w_2}\bar{x}_1 + \frac{w_2}{w_1 + w_2}\bar{x}_2$$

$$s^2 = \frac{w_1}{w_1 + w_2}s_1^2 + \frac{w_2}{w_1 + w_2}s_2^2 + \left(\frac{w_1}{w_1 + w_2}\right)\left(\frac{w_2}{w_1 + w_2}\right)(\bar{x}_1 - \bar{x}_2)^2$$

ここまでは，同じデータを1クラスターと見たときと，2クラスターと見たときの標本統計量の関係に過ぎない。ところが，「積率法」の名の通り，標本統計量と，それに対応するパラメターを等しいと置けば，1クラスターのパラメターと2クラスターのパラメターの間に近似的に成り立つ等式として，読むことができる。即ち，$\bar{x} = \mu$，$\bar{x}_i = \mu_i$，$s = \sigma$，$s_i = \sigma_i$と置く。この式が，まさにSSの公式なのである。この解釈に基づけば，2次元への拡張の指針は明快である。融合すべき2つのコンポーネントを，ウェイトw_1を持った$N\left(\mu_{1x}, \mu_{1y}, \sigma^2_{1x}, \sigma^2_{1y}, \rho_1\right)$とウェイト$w_2$を持った$N\left(\mu_{2x}, \mu_{2y}, \sigma^2_{2x}, \sigma^2_{2y}, \rho_2\right)$としよう。適切に融合された分布は，ウェイト$w$を持つ

$N(\mu_x,\mu_y,\sigma_x^2,\sigma_y^2,\rho)$であるとする。4つのパラメターとウェイトは1次元の公式から借りて来る。

$$\mu_x = \frac{w_1}{w_1+w_2}\mu_{1x} + \frac{w_2}{w_1+w_2}\mu_{2x}$$

$$\sigma_x^2 = \frac{w_1}{w_1+w_2}\sigma_{1x}^2 + \frac{w_2}{w_1+w_2}\sigma_{2x}^2 + \left(\frac{w_1}{w_1+w_2}\right)\left(\frac{w_2}{w_1+w_2}\right)(\mu_{1x}-\mu_{2x})^2$$

$$\mu_y = \frac{w_1}{w_1+w_2}\mu_{1y} + \frac{w_2}{w_1+w_2}\mu_{2y}$$

$$\sigma_y^2 = \frac{w_1}{w_1+w_2}\sigma_{1y}^2 + \frac{w_2}{w_1+w_2}\sigma_{2y}^2 + \left(\frac{w_1}{w_1+w_2}\right)\left(\frac{w_2}{w_1+w_2}\right)(\mu_{1y}-\mu_{2y})^2$$

$$w = w_1 + w_2$$

問題は，どうやってρを決定するか，である。2つのコンポーネントから，合わせてサイズnのサンプルを生成したとしよう。それぞれのウェイトを考慮に入れれば，$n_1 = \frac{w_1}{w_1+w_2}n$の観測値は，第1の分布から，$n_2 = \frac{w_2}{w_1+w_2}n$の観測値は第2の分布から来る。観測値の番号を調整し，$1,\cdots\cdots,n_1$は第1のサンプル，$n_1+1,\cdots\cdots,n_1+n_2$は第2のサンプルとする。我々の議論が，1次元の場合と平行になるように，以下の文字を導入する。

X_i	Y_i	第i観測値のx, y座標
$\bar{x}_1$	$\bar{y}_1$	第1のサンプルの平均
$\bar{x}_2$	$\bar{y}_2$	第2のサンプルの平均
$\bar{X}$	$\bar{Y}$	融合されたサンプルの平均

そこで，第1のサンプル，第2のサンプル，融合されたサンプルの標本相関係数は，

$$R_1 = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1} X_i Y_i - \bar{x}_1 \bar{y}_1}{\sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}(X_i - \bar{x}_1)^2}\sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}(Y_i - \bar{y}_1)^2}} = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1} X_i Y_i - \bar{x}_1 \bar{y}_1}{S_{1x} \cdot S_{1y}}$$

$$R_2 = \frac{\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} X_i Y_i - \bar{x}_2 \bar{y}_2}{\sqrt{\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}(X_i - \bar{x}_2)^2}\sqrt{\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}(Y_i - \bar{y}_2)^2}} = \frac{\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} X_i Y_i - \bar{x}_2 \bar{y}_2}{S_{2x} \cdot S_{2y}}$$

$$R = \frac{\frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2} X_i Y_i - \overline{X}\,\overline{Y}}{\sqrt{\frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2}(X_i - \overline{X})^2}\sqrt{\frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2}(Y_i - \overline{Y})^2}} = \frac{\frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2} X_i Y_i - \overline{X}\,\overline{Y}}{S_x \cdot S_y}$$

R_1 と R_2 から,

$$\sum_{i=1}^{n_1} X_i Y_i = n_1\left(R_1 \cdot S_{1x} \cdot S_{1y} + \bar{x}_1 \bar{y}_1\right)$$

$$\sum_{i=n_1+1}^{n_1+n_2} X_i Y_i = n_2\left(R_2 \cdot S_{2x} \cdot S_{2y} + \bar{x}_2 \bar{y}_2\right)$$

そこで R の分子の一項目は,

$$\frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2} X_i Y_i = \frac{1}{n_1+n_2}\left\{n_1\left(R_1 \cdot S_{1x} \cdot S_{1y} + \bar{x}_1 \bar{y}_1\right) + n_2\left(R_2 \cdot S_{2x} \cdot S_{2y} + \bar{x}_2 \bar{y}_2\right)\right\}$$

$$\therefore \quad R = \frac{\frac{1}{n_1+n_2}\left\{n_1\left(R_1 \cdot S_{1x} \cdot S_{1y} + \bar{x}_1 \bar{y}_1\right) + n_2\left(R_2 \cdot S_{2x} \cdot S_{2y} + \bar{x}_2 \bar{y}_2\right)\right\} - \overline{X}\,\overline{Y}}{S_x \cdot S_y}$$

n_1, n_2 を w_1, w_2 による表現で置き換え,

$$R=\frac{\dfrac{1}{w_1+w_2}\left\{w_1\left(R_1\cdot S_{1x}\cdot S_{1y}+\bar{x}_1\bar{y}_1\right)+w_2\left(R_2\cdot S_{2x}\cdot S_{2y}+\bar{x}_2\bar{y}_2\right)\right\}-\overline{XY}}{S_x\cdot S_y}$$

最後に，

$$\bar{x}_1=\mu_{1x}\qquad \bar{x}_2=\mu_{2x}$$

$$\bar{y}_1=\mu_{1y}\qquad \bar{y}_2=\mu_{2y}$$

$$s_{1x}=\sigma_{1x}\qquad s_{2x}=\sigma_{2x}$$

$$s_{1y}=\sigma_{1y}\qquad s_{2y}=\sigma_{2y}$$

$$R_1=\rho_1\qquad R_2=\rho_2$$

$$S_x=\sigma_x\qquad S_y=\sigma_y\qquad R=\rho$$

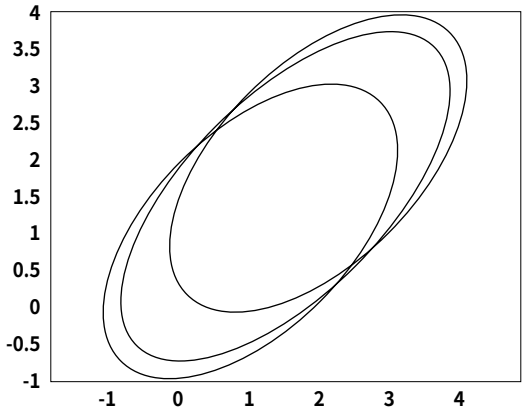
と置いて，

$$\rho=\frac{\dfrac{1}{w_1+w_2}\left\{w_1\left(\rho_1\cdot\sigma_{1x}\cdot\sigma_{1y}+\mu_{1x}\mu_{1y}\right)+w_2\left(\rho_2\cdot\sigma_{2x}\cdot\sigma_{2y}+\mu_{2x}\mu_{2y}\right)\right\}-\mu_x\mu_y}{\sigma_x\sigma_y}$$

を得る。

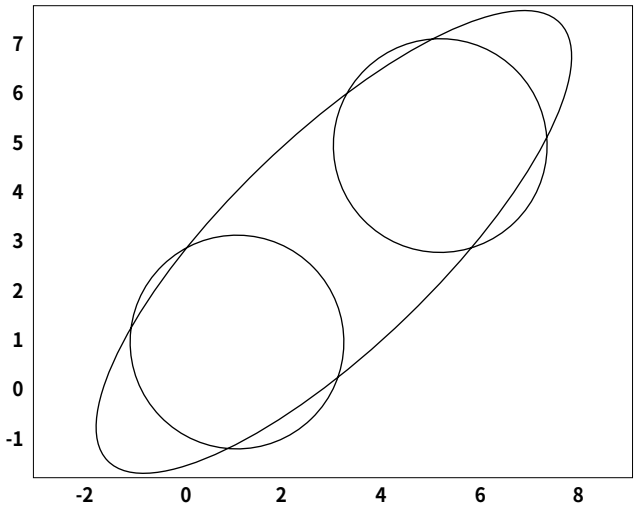
この方法によって，どんな融合が生じるか，視覚的に確かめるために，2つの例を挙げる。次の図には，3つの楕円が描かれている。

図4



中央右上寄りの楕円は，パラメター $\mu_x = 2$ $\mu_y = 2$ $\sigma_x = 1$ $\sigma_y = 1$ $\rho = 0.5$ の2次元正規分布の確率90パーセントの等高線である。中央左下寄りの楕円は，パラメター $\mu_x = 1$ $\mu_y = 1$ $\sigma_x = 1$ $\sigma_y = 1$ $\rho = 0.5$ の同じ確率90パーセントの等高線である。両者には等しいウェイトが与えられており，相似測度は0.717である。二つにまたがって描かれている，大きな楕円が，融合された1クラスター・モデルである。パラメターは $\mu_x = 1.5$ $\mu_y = 1.5$ $\sigma_x = 1.118$ $\sigma_y = 1.118$ $\rho = 0.6$ である。図5では，それぞれ $\rho = 0$ でx,y 方向の分散の等しい，2つの2次元正規分布を融合させることを，試みる。

図5



しかし，図からわかるように，90％の等高線は遠く離れている。相似測度はほとんど0である。明らかに，これは，KRの実際の降下プロセスで，融合させるべきケースではない。それにもかかわらず，我々の公式は適度に二つの「中を取って」いる。楕円は二つの円の面積の90％近くをカバーし，それらの位置関係を正しくとらえ，しかも大き過ぎない。

8．2次元のための Phase3（L2E+MEASURE2）

MEASURE2の検討から始めよう。5章では，1次元のMEASURE2の具体的な形にまで，説き及ばなかったが，1次元と2次元は形式的に類似しているので，ここでは2次元のみ扱う。*sim*2のf_1をλ，f_2をγと書き換えたのが，次の式である。

$$sim3(\lambda,\gamma) = \frac{\iint \lambda(x,y)\gamma(x,y)dydx}{\sqrt{\iint \lambda^2(x,y)dydx}\sqrt{\iint \gamma^2(x,y)dydx}}$$

λをnコンポーネント・モデル

$$\lambda = p_1 f_1 + p_2 f_2 + \cdots\cdots\cdots\cdots + p_n f_n$$

γ を（n-1）コンポーネント・モデル

$$\gamma = q_1 g_1 + q_2 g_2 + \cdots\cdots\cdots + q_{n-1} g_{n-1}$$

とする。各コンポーネントのパラメターは $f_i \sim N(\mu_i, \Sigma_i)$ $g_j \sim N(\nu_j, \Omega_j)$ と書く。
*sim*3中の3つの積分を求めてみよう。

$$\lambda\gamma = \sum_{i=1}^{n}\sum_{j=1}^{n-1} p_i f_i q_j g_j = \sum_{i=1}^{n}\sum_{j=1}^{n-1} p_i q_j f_i g_j$$

$$\therefore \quad \iint \lambda(x,y)\gamma(x,y)dydx = \sum_{i=1}^{n}\sum_{j=1}^{n-1} p_i q_j \iint f_i(x,y) g_j(x,y) dydx$$

6章の（1）を用いて,

$$= \sum_{i=1}^{n}\sum_{j=1}^{n-1} p_i q_j \phi_{\Sigma_i+\Omega_j}(\mu_i - \nu_j)$$

$$\begin{aligned}\lambda^2 &= (p_1 f_1 + p_2 f_2 + \cdots\cdots\cdots + p_n f_n)^2 \\ &= p_1^2 f_1^2 + p_2^2 f_2^2 + \cdots\cdots\cdots + p_n^2 f_n^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_i f_i p_j f_j \\ &= \sum_{i=1}^{n} p_i^2 f_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_i p_j f_i f_j\end{aligned}$$

$$\therefore \quad \iint \lambda^2(x,y)dydx = \sum_{i=1}^{n} p_i^2 \iint f_i^2(x,y)dydx + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_i p_j \iint f_i(x,y) f_j(x,y)dydx$$

（1）を用いて,

$$= \sum_{i=1}^{n} p_i^2 \phi_{2\Sigma_i}(0) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_i p_j \phi_{\Sigma_i+\Sigma_j}(\mu_i - \mu_j)$$

$$\gamma^2 = \left(q_1 g_1 + q_2 g_2 + \cdots\cdots\cdots + q_{n-1} g_{n-1}\right)^2$$

$$= q_1^2 g_1^2 + q_2^2 g_2^2 + \cdots\cdots\cdots + q_{n-1}^2 g_{n-1}^2 + 2\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} q_i g_i q_j g_j$$

$$= \sum_{i=1}^{n-1} q_i^2 g_i^2 + 2\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} q_i q_j g_i g_j$$

$$\therefore \quad \iint \gamma^2(x,y)dydx = \sum_{i=1}^{n-1} q_i^2 \iint g_i^2(x,y)dydx + 2\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} q_i q_j \iint g_i(x,y) g_j(x,y)dydx$$

（1）を用いて

$$= \sum_{i=1}^{n-1} q_i^2 \phi_{2\Omega_i}(0) + 2\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} q_i q_j \phi_{\Omega_i+\Omega_j}\left(\nu_i - \nu_j\right)$$

L2Eはパラメター推定の新しい方法である。fを真の確率分布，$\hat{f}$ をその推定量とするとき，

$$\int\left\{\hat{f}(x;\theta) - f(x)\right\}^2 dx$$

を最小化することを考える。積分記号は，確率変数xの定義域を全てカバーする定積分である。次元については，xの次元に応じて，何次元の積分と考えても良い。もちろん，この形は，直接取り扱えない。なぜならば，そこにある真の確率分布は，わかっていないからである。しかし，二乗の項を展開して，

$$\int \hat{f}(x;\theta)^2 dx - 2\int \hat{f}(x;\theta) f(x)dx + \int f^2(x)dx$$

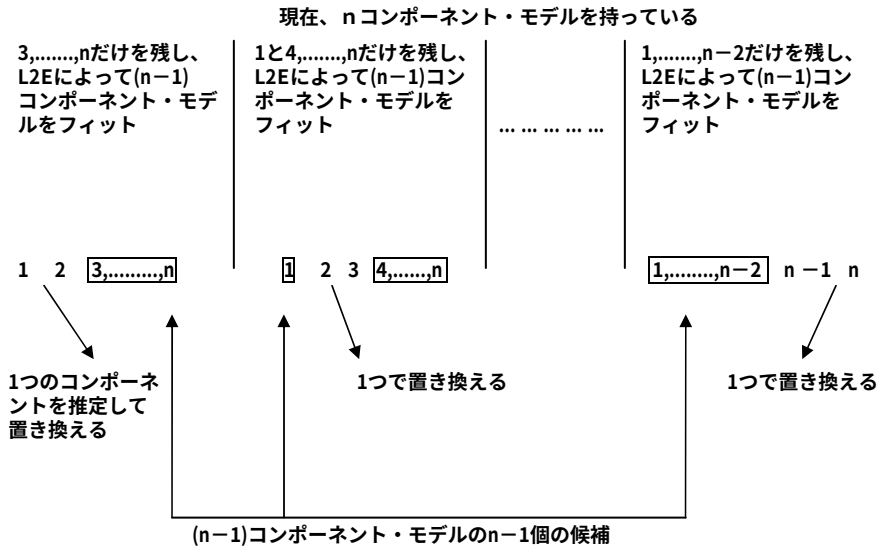
最後の項は，推定すべきパラメターを含まないので，最適化から取り除ける。二番目の積分は $E\hat{f}(x;\theta)$ だから，$\frac{1}{N}\sum_{i=1}^{N} \hat{f}(x_i;\theta)$ によって推定できる。Nはデータ・サイズ，x_i は観測値である。最初の積分に閉じた形があるということは，同様の計算を繰り返しやって来たのでわかるだろう。そこで，L2Eは

$$\hat{\theta} = \arg\min_{\theta}\left[\int \hat{f}(x;\theta)^2 dx - \frac{2}{N}\sum_{i=1}^{N} \hat{f}(x_i;\theta)\right]$$

と定式化される。Phase3は，コンポーネントの数が，かなり少なくなったとこらから始まるが，この定式化は，まだたくさんのパラメターを含んでいる。より信頼できるパラメターを得るために，L2Eで全てのコンポーネントを推定するのではなく，部分的に推定する方法を考えよう。

そのために，図6を見ながら，5章で述べたPhase3の考え方を復習する。説明を簡単にするために，1次元に戻ろう。今，nコンポーネント・モデルから，n-1コンポーネント・モデルに降下するところだと想定する。さらに，n個のコンポーネントは，それぞれの平均の大きさの順に，左から右へと番号がついていることにする。融合は隣り合ったコンポーネント同士について，考える。縦棒でしきった列を左から右に見ていくことにし，我々は，左端の2つのコンポーネントの融合から出発する。nコンポーネント・モデル中の3,……,nを残し，コンポーネント1と2とを置き換えるべき，1つのコンポーネントのパラメターを推定する。次に，これと3,……,nを合わせたn-1コンポーネント・モデルと元のnコンポーネント・モデルとの間の相似測度を計算する。第二の場合に移り，nコンポーネント・モデルから，1,4,……,nを残したまま，コンポーネント2, 3を置き換えるべき，1つのコンポーネントのパラメターを推定する。それから，この新しいn-1コンポーネント・モデルとnコンポーネント・モデルの間の相似測度を計算する。このように続けて行って，右端の2つのコンポーネントを融合するケースにまで達する。そこで，我々はn-1個のn-1コンポーネント・モデルを候補として持つことになる。これらから，nコンポーネント・モデルとの相似測度が最大のものを選択する。

図６



1次元で，SSが融合を試みたのは，二種類のペアの取り方のみである。即ち，隣どうしのペアと「一つ飛び」のペア（“two-away”, ペアの相手が，もう一つ別なコンポーネントを飛び越した，向こう側にある場合）である。もちろん，2次元においては，1次元のような順序構造は存在しない。そこで，原始的なサーチ・アルゴリズムを組むなら，$\binom{n}{2}$個の全ての融合を試みなければならない。仮に，何等かの「近傍概念」によって，融合に値するペアを判別できる

なら，そのような判別に残ったペアのみを，融合させることが望ましい。しかし，計算の手間がかからず，荒っぽい予備選考にならない，そのような概念を見つけることも難しいので，本研究では，原始的なサーチを実行した。

Phase3では，最適なクラスター数選択のために，各nにおいて選ばれたモデルの情報量基準を計算する。我々のアルゴリズムではAICとBICを計算し，そのモデル選択における得失については，次章で検討する。既に予告した通り，MEASURE1と2の違いは重要である。しかし，これはシミュレーション結果を見ることによって，最も効率的に説明できる。そこで，次章の後に独立の章を設けて議論する。